秋田大学教育文化学部研究紀要
教育科学第75集別刷　令和２年３月

Revision of Examinations in a 1st Year University English Program: The practicality and reliability of using in-house versus textbook publishers' examinations

Adrian Paterson & William Bradley Horton

Akita University

秋田大学教育文化学部研究紀要　教育科学部門　75　pp. 1～5　2020

Revision of Examinations in a 1st Year University English Program: The practicality and reliability of using in-house versus textbook publishers' examinations

Adrian Paterson & William Bradley Horton

Akita University

Introduction

This study is intended to evaluate the prospects for changes to exam creation procedure and grading in a multi-class English course as part of an effort to create a more efficient, yet still effective educational environment. Simultaneously, it is hoped that this brief study will contribute to rational, objective decision making in language teaching by providing an example of exam evaluation.

Need for time efficiency

Incessant demand for change in Japanese universities in recent years, motivated in part by repeated reduction in funding and increases in expectations by MEXT, have come to a critical point, thanks to pressure for faculty to conform to Japanese labor law limitations of 38.5 hours per week. Exemption of faculty from these laws had allowed Japanese universities to largely keep their administrative staff within the legally allowed labor hours by shifting administrative tasks to faculty, and blurring lines of responsibility between faculty (researchers and teachers, but also involved in policy decisions) and staff (essential support staff for all university activities and implementation of policies). With these new requirements for faculty to conform to the same labor laws as the rest of society, time commitment of faculty to administration and teaching must be drastically reduced, while not merely maintaining the same quality of instruction, but if possible improving opportunities for students.

Simplification and streamlining of administrative procedure is beyond the ken of the authors, and the absolutely essential coordinated changes to the curriculum needed in all faculties of the university require time and consensus, and probably the acquiescence of MEXT. It is thus essential for us to consider incremental changes within the existing curriculum which can contribute to making Akita University an effective university while trying to reduce aggregate faculty time commitments for required teaching, research, and other necessary administrative tasks to the legal limits. Naturally, many of the same considerations are valid at other universities.

Targeting examinations

The production of tests requires a complex, time consuming and often highly technical design cycle of identifying test purpose, setting criteria, defining the construct, item design, evaluating and prototyping, final formatting, analysis, and decision making (Fulcher, 2010). This is often beyond the time constraints of most university instructors, and thus impossible to carry out effectively. Bachman and Palmer (1996, 2010) define the usefulness of a test in terms of its validity, practicality, and reliability. For a test to be valid, it should demonstrably measure the construct it purports to measure. The practicality of a test is the amount of resources in terms of people-time and money required to produce, administer, and grade it. For a test to be reliable it should ensure "that responses to individual items are not dependent upon the responses to other items, that they have good facility values and discrimination, and that we have enough items" (Fulcher & Davidson, 2007, p.104). This study aims to maximize the practicality of a course test without unnecessarily sacrificing validity and especially reliability. Careful design is particularly important for multichoice tests, where poorly designed items can increase the possibility of test takers guessing the correct answer rather than using linguistic knowledge, or make it nearly impossible to choose the correct option (Hughes, 2003).

Examination of current faculty time commitments in the English for Academic Purposes (大学英語) I & II first year courses indicates that exam preparation and grading are two areas in which time commitments could potentially be reduced. Currently, two faculty members are assigned as a team to create each exam, with equally weighted exams given twice each semester, covering the previous seven weeks of material. While the textbook used remains the same for a number of years, the chapters covered changes from year to year, ensuring that students cannot obtain previous exams, and producing a range of different English language knowledge within the student body, while hopefully allowing instructors to bring "fresh" material to the classroom. However, this also means that four times each year, two teachers need to design a new exam, check it for errors, circulate it to all English instructors for comments, and revise the exams again. Even a reduction of 1 hour per exam designer would result in more than 1 full day of professorial time per year (1 hour x 2 exam writers x 4 exams), and that does not include any potential reduction of time by other English instructors. If greater efficiency in exam creation could be achieved, it seems self-evident that it would more likely total much more than 1 professorial-work day.

Based on these observations, we determined it would be best to try to see if there was exam software which could provide a quicker alternative to our current exam making procedures. An exam making program from the previous edition of the textbook was available, and we decided to provide an exam as a "practice exam" for students, with the hope that exam results would indicate whether this "practice exam" was a fair equivalent, or perhaps a conceivable alternative to designing an exam from scratch.

Method

Course Exam structure

Each of the four examinations for the first year English courses covers material from one chapter of the current textbook, as well as other materials covered during those seven weeks. Each exam consists of six sections, including sections on the two readings and an unseen reading selected by one of the instructors responsible for creation of the exam. These sections cover approximately 50% of each exam. Other sections cover the video and common material for all courses related to conversation and writing—largely outside the purview of the textbook. Each exam is thus basically new, generally covering different chapters from the textbook each year. The other sections include a listening section using the video from each unit from the DVD that comes with the teacher's pack for the textbook. A writing section that usually tests knowledge of structural aspects of academic writing, such as introductions, conclusions, thesis statements, and topic sentences. The last section is a conversation section, which is based on model conversations from handouts provided to students for group work in class. These model conversations are usually tested by asking students to complete deleted sections with sentences and phrases from the original conversation. By including them in a written exam like this, it becomes an exercise in memorization; in fact, the most successful students getting perfect or near perfect scores are those who memorize the script, whereas students who rely on their linguistic sense or skills tend to lose points. The design thus forces students to rely on such rote learning strategies at the expense of important conversation skills such as discourse knowledge, schema activation, and pronunciation, the neglect of which is particularly problematic (Paterson, 2018).

Practice exam

A practice exam consisting of the entire set of 30 questions available on the previous edition of the National Geographic Reading Explorer exam software package was created by the researchers. Examination of the questions indicated that they covered the same material as in the current edition of the textbook. In order to eliminate possible biases by the researchers, the entire set of questions was then used. These included questions related to vocabulary and content of the two readings, as well as a related unseen reading for testing reading ability. Students were instructed to write their answers on both the question sheet and the answer sheet. At the end of the practice exam, the answer sheets and consent forms were collected, and then students were given the correct answers, which they self-marked on the question sheets. This meant that students had immediate feedback on the practice exam.

Subjects

Four first-year English classes taught by three instructors were selected. The data collected was anonymized by randomly assigning each class a letter code; class A was an advanced class of education and humanities majors, classes B and C were advanced classes of engineering majors, and class D was an intermediate class of health majors. The terms 'advanced' and 'intermediate' in this context are used to indicate the students level of ability relative to other classes in the course, assigned through an initial placement exam, and do not relate directly to students' actual level of ability as measured by any external standards. Once the two exams were matched, numeric codes were randomly assigned and identifying personal information was deleted.

The practice exam was given to students in all four classes the week before the exam, thus before most students had begun to prepare for the exam, but after completion of the assignments. The purpose of the test was explained, and consent for use of data obtained in writing before the exam[1]. Students were given 35 minutes to complete the exam which seemed to be sufficient, though less than half the regular exam time. The practice exam papers were graded by hand and checked by computer during data entry to ensure accuracy.

There were a total of 110 subjects who were present for the practice exams and who gave their written consent to use exam results. However, the researchers were unable to obtain the course exams for classes C and D in time to analyze the results for this study due to overzealous enforcement of deadlines, so only regular course exam results for classes A and B were used for comparison of the two exams and for item analysis and reliability analysis of the course exam. The practice

[1] Consent for five subjects was not obtained (one checked 'no' and four did not submit the consent form). Exam data from those individuals has not been utilized here.

exam results were analyzed for all four classes.

Results and Discussion

The answer sheets from the practice exam were all marked manually by the second author, and then the students' raw answers were entered into a spreadsheet for analysis by the first author. All of the questions were multichoice, and the spreadsheet was set up to score the answers. This was done as a double check to ensure the accuracy of the results. The data was then imported into SPSS for analysis. Descriptive statistics are shown below in Table 1.

Table 1. Practice exam descriptive statistics

Class	N	Range (min-max)	Mean	Standard Deviation
A	31	16-27	21.48	2.719
B	24	10-25	18.29	4.048
C	24	16-28	22.92	3.269
D	31	10-25	18.03	4.086

The results of the course exams were hand-marked by each class teacher and then students' raw answers for multichoice questions and teachers' evaluations of long answer questions were entered into a spreadsheet for analysis. The spreadsheet was set up to grade the multichoice answers and total the scores. This acted as a double-check to ensure accuracy of results. The data was then imported into SPSS for analysis. Descriptive statistics are below in Table 2.

Table 2. Course exam descriptive statistics

Class	N	Range (min-max)	Mean	Standard Deviation
A	31	73-98	84.48	6.239
B	24	61-92	79.33	7.982

Comparison of the two tests

There was a significant but weak correlation between the practice exam and the course exam (Pearson r = 0.387, N = 55, p <0.01) for classes A and B. This means that only 14.97% of the variance of scores on the course exam can be explained by the score on the practice exam. However, in an informal survey of class B after the mid-semester exam, about half of the students indicated by show of hands that they had been disappointed with their results on the practice test, and so they had studied harder than they would have normally for the mid-semester exam. None indicated that they had studied less because they had gotten a high score on the practice exam, but that may be due to so few of them getting such high scores on the practice exam. However, this motivating or demotivating effect could help to explain this lack of consistency between the two exams.

Another explanation for this weak correlation, could be due to the fact that the practice test was only based on the course textbook, whereas the course exam had sections that were not related to the textbook, but rather to other materials taught in the course. However, comparisons between the reading subsection of the practice exam and the textbook related video and reading sections on the course exam all resulted in similarly low correlations, none of which were statistically significant. These are summarized in Table 3.

Table 3. Correlations between reading subsection of the practice exam and textbook related sections of the course exam

Course exam Section	Pearson *r*	N	Sig. (2-tailed)
I. Video	0.070	55	0.609
II. Reading A	0.119	55	0.386
III. Reading B	0.137	55	0.318

Item Analysis

An analysis of the internal reliability of both tests exams was carried out using SPSS. This measures "the degree to which individual items or groups of items on a test correlate with tone another" (Davies et al., 1999, p.86). A test with a Cronbach's alpha less than .7 is usually considered to have low internal reliability (Pallant, 2005; Salkind, 2008). The practice exam had a Cronbach's alpha of 0.732 (N of items = 30), and the course exam had a Cronbach's alpha of 0.580 (N of items = 48). Generally speaking, one way to improve the internal reliability of a test is to increase the number of items, therefore, the fact that the course exam with more items has a much lower alpha value suggests that the items in the practice test are more consistent. However, as the course exam has sections not directly related to the textbook which test other language skills, it is more likely that internal consistency will be lower.

Following analysis of internal reliability, item analyses on both tests were carried out using a copy of the scoring spreadsheets. Item facilities (IF) and item differentiations (ID) were calculated for each item. Brown (2005, p.75) suggests that acceptable items should have an IF between .30 and .70, and an ID greater than .30, items with an ID between .20 and .29 should be modified, and items with an ID less than .19 should be discarded.

Of the 30 items in the practice exam, 11 had an IF within the acceptable limits, three were lower than .30, and 16 were greater than .70. Four of the five lowest item facilities on the practice exam were for items 11, 12, 13, and 14. The reason for this was that the format of these questions required students to choose multiple options when all other questions in the test required a single option, apparently causing confusion to most students. The IDs for all of these questions were positive, meaning that while they were more difficult, they affected top scoring students less than lower scoring students. It should be noted that the more recent version of the test making software eliminates this type of item and adopts the more common, and less confusing, single answer multichoice item format. All of the 30 items on the practice exam were positive, 13 items had an ID greater than .30, four had an ID between .20 and .29, and the remaining 13 had an ID below .19. Of the very low ones, eight were for items with an IF greater than .89, meaning that they were too easy for the subjects of this study and that even the lowest scoring students could answer them. It should be noted that if this exam were to be adopted course-wide, it would need to be possible for students in lower level classes to pass it, so it is desirable for some items to be too easy for the top students.

The course exam had 48 items, of which 11 had an IF within the acceptable range, the rest were greater than .70 and four of them were 1.00 meaning that they were too easy for all students and therefore did not contribute to differentiating levels of ability. Of greater concern is the fact that only 14 items had an ID greater than .20, of the remainder five were zero and three were negative, meaning that they either do not discriminate ability levels at all or lower level students actually performed better on them.

Conclusion

Making an effective test requires a lot of skill and effort. The in-house course exam in this study probably required two faculty members to spend 10 to 15 hours each producing it, and then the remaining teachers would have needed to spend an average of 1 to 2 hours each proofreading it. In the current university environment, in which faculty members are required to carry out many administrative tasks outside of those directly related to teaching and research. It is quite understandable that few have enough time to put in the effort required to help produce effective course exams, instead leaving it to others to check them. Such a system can easily fall apart when nobody has the time to carry out the vital task of proofreading. Therefore, even merely considering education (and not the legal time constraints or the demand for research time) we feel that it is imperative to find ways to reduce the burden on teachers. Textbook publishers' test making software offers one alternative to our current inefficient exam making procedure. One of the greatest benefits to this altenative approach to course test making is that the test items have already been trialed and therefore do not require such a significant number of instructor-hours to produce an effective test. The practice test exam used in this study took approximately 30 minutes for the second author to produce and proofread. This is insignificant compared to the 30 to 50 manhours required to produce the in-house course exam.

The situation is further exacerbated by the inability (due to lack of time and training) of teachers to do the kind of follow up analysis that we did in this study. It is only by analyzing the results of a test using statistical methods that teachers can identify items that perform well and to learn from their mistakes. This may actually be the first attempt to conduct such an evaluation for this course. When using a textbook publisher's test making software, generally speaking, these bugs have already been ironed out, reducing the need for such follow-up analysis.

One limitation of this study was that the classes that participated in the study all had students with higher ability levels. A larger study with a wider range of ability levels would likely have produced quite different results. In particular, it would have reduced the tendency for items to have such high item facilities.

The textbook publisher's test was also administered as a practice test before most students had begun their exam preparation, therefore, they were not as well prepared for it as they were for the course exam. This was probably one of the factors contributing to the low correlation between the two exams. It is quite likely that had students prepared for the practice test, the results may have been quite different resulting in a stronger correlation.

Unfortunately, we did not obtain a copy of the updated Exam View software in time to use in this study. A preliminary review of the complete set of available items showed that it has a much larger pool of potential readings and questions, including four reading sections as opposed to one, and a larger pool of vocabulary items, 30 items as opposed to 19 items in the previous version. This included dropping the four problematic multi-answer multichoice questions in favor of the more conventional single-answer multichoice format. Had the current software version been available in time to be used in this study we may have seen somewhat different results. However, administering it again as a practice exam rather than a course exam could mean that students would be similarly unprepared for it, and thus a second round of testing seemed unnecessary.

In conclusion, using test making software provided by the textbook publisher to produce exams may only offer a small advantage over in-house exams in terms of reliability, however, in terms of practicality they offer significant benefits for busy teachers. They take considerably less time to produce, and because they have already been trialed and refined by the authors and publishers, they do not require such careful proofreading by all teachers involved. Another big advantage is that they can be produced in multiple versions, each with its own unique answer key, which is important when different classes take the exam at different times.

Bibliography

Bachman, L. F., & A. S. Palmer. (1996). *Language testing in practice: Designing and developing useful language* *tests*. Oxford: Oxford University Press.

Bachman, L. F., & A. S. Palmer. (2010). *Language Assessment in Practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment.* New York: McGraw-Hill.

Davies, A., A. Brown, C. Elder, K. Hill, T. Lumley, & T. F. McNamara. (1999). *Dictionary of language testing.* Cambridge: Cambridge University Press.

Douglas, N. & D. Bohlke. (2015). *Reading Explorer 3* (2nd ed.). Boston: National Geographic Learning/ Cengage Learning.

Fulcher, G. (2010). *Practical language testing.* London, UK: Hodder Education.

Fulcher, G., & F. Davidson. (2007). *Language testing and assessment: An advanced resource book*. London: Routledge.

Hughes, A. (2003). *Testing for Language Teachers* (2nd ed.). Cambridge, UK: Cambridge University Press.

Pallant, J. (2005). *SPSS survival manual: A step by step guide to data analysis using SPSS for Windows (Version 12)* ([Rev.]). Crows Nest, N.S.W.: Allen & Unwin.

Paterson, A. (2018). "'Sensei. How do I say this?' A case for required pronunciation courses for English teacher trainees in Japan," *Journal of Anglo-American Studies* no. 42: 33-62.

Salkind, N. J. (2008). *Statistics for people who (think they) hate statistics* (3rd ed.). Los Angeles: Sage.

Software

Reading Explorer (Nancy Douglas/Paul MacIntyre/John Chapman). ExamView Assessment Suite. (2010). Heinle Cengage Learning.

Reading Explorer 1-3, Second Edition, ExamView. (2015). National Geographic Learning/Cengage Learning.